EURASIP Journal on Applied Signal Processing 2005:9, 1334–1349

Anthropomorphic Coding of Speech and Audio: A Model Inversion Approach

Christian Feldbauer
Signal Processing and Speech Communication Laboratory, Graz University of Technology, 8010 Graz, Austria
Email: feldbauer@tugraz.at

Gernot Kubin
Signal Processing and Speech Communication Laboratory, Graz University of Technology, 8010 Graz, Austria
Email: g.kubin@ieee.org

W. Bastiaan Kleijn
Department for Signals, Sensors and Systems, KTH (Royal Institute of Technology), 10044 Stockholm, Sweden
Email: bastiaan.kleijn@s3.kth.se

Received 14 November 2003; Revised 25 August 2004

Auditory modeling is a well-established methodology that provides insight into human perception and that facilitates the extraction of signal features that are most relevant to the listener. The aim of this paper is to provide a tutorial on perceptual speech and audio coding using an invertible auditory model. In this approach, the audio signal is converted into an auditory representation using an invertible auditory model. The auditory representation is quantized and coded. Upon decoding, it is then transformed back into the acoustic domain. This transformation converts a complex distortion criterion into a simple one, thus facilitating quantization with low complexity. We briefly review past work on auditory models and describe in more detail the components of our invertible model and its inversion procedure, that is, the method to reconstruct the signal from the output of the auditory model. We summarize attempts to use the auditory representation for low-bit-rate coding. Our approach also allows the exploitation of the inherent redundancy of the human auditory system for the purpose of multiple description (joint source-channel) coding.

Keywords and phrases: speech and audio coding, auditory representation, auditory model inversion, auditory synthesis, perceptual domain coding, multiple description coding.

1. INTRODUCTION

1.1. Motivation

The encoding of an analog signal at a finite rate requires quantization and introduces distortion. Models of the human auditory system can be exploited to minimize, for a given rate (specified either as an average or as a fixed rate), the audible distortion (as quantified by the model) introduced by the encoding [1, 2, 3]. Signal features will then be specified with a precision that reflects audible distortion. However, the introduction of knowledge of the auditory system into coding has been handicapped by delay and computational constraints. For instance, temporal masking and the adaptation of the hearing system to a stimulus are highly nonlinear effects [4, 5]. A time-localized quantization error in the perceived signal can result in a significant change in the auditory nerve firings over a response time interval that can last on the order of hundreds of milliseconds. Therefore, the effect of time-localized quantization errors that are hundreds of milliseconds apart cannot be separated into additive terms. As a result, it is difficult to include such dependencies of quantization errors during the quantization process.

The simple distortion criteria used in practical systems result from a desire to perform efficient quantization at reasonable computational complexity. Such efficient, low-complexity quantization is facilitated by three conditions: (i) the (vector) variable is of low dimension, (ii) the distortion criterion is a single-letter one (i.e., the distortion measure is a sum over many sample distortions), and (iii) the variables are independent. This is particularly well illustrated by the discrete-cosine-transform (DCT) -based lapped

transforms commonly used in audio coding [6]. These transforms allow a spectrally weighted mean-square error distortion measure to be approximated as a single-letter criterion. For wide-sense stationary signals, the results of the DCT are asymptotically equivalent to the results of the Karhunen-Loève transform, thus performing an approximate decorrelation of the data. Finally, scalar quantization is used to have low complexity.

Our objective is to use sophisticated auditory-model-based distortion criteria without the significant approximations commonly used (such as simple error-weighting filters in linear-prediction-based speech coders or the exclusive consideration of frequency-domain masking in many audio coders).

Most quantitative models of the human auditory perception provide an auditory representation of the acoustic signal as output. However, the models generally do not include a quantitative measure of the perceptual distance of two realizations of the auditory representation. In [7], a correlation measure of the internal representations was proposed as an objective distortion measure. Such a measure is closely related to a single-letter weighted squared-error measure. We will assume that a single-letter distortion criterion on the auditory representation can provide a high-quality distortion measure.

The usage of sophisticated distortion criteria within the existing coding architectures leads to so-called delayed-decision coding. Delayed-decision coding methods have been used in the context of a squared-error criterion and linear-prediction-based waveform coding (e.g., [8]). In the delayed-decision approach, the quantization of a signal block is decided only after consideration of the quantization of a certain number of future blocks. Even when using pruning procedures that eliminate the consideration of unlikely configurations, this method becomes computationally very expensive for distortion measures that have the long time responses associated with hearing models [9]. This motivates the consideration of less conventional coding architectures.

The coding approach we presented in [10], which is the basis throughout this paper, avoids the high computational complexity of the delayed-decision approach by exploiting the single-letter nature of the criterion in the auditory representation. The signal is transformed to the auditory domain and coded in that domain. The decoding is followed by a transform back towards the acoustic domain. The transform from the acoustic to the auditory domain can be many-to-one, making the inverse transform in general nonunique. This auditory-domain approach towards coding allows the usage of a single-letter distortion criterion and yet accounts for the dependency of perceived distortion on errors in the signal that are far apart in time.

It is important to note that virtually all state-of-the-art speech and audio coding methods operate on a block-by-block basis (e.g., [1, 2, 6, 8]). For subband/transform coding for example, decimated filterbanks or lapped transforms are used, which introduce block boundaries at regular time positions (often even independent of the actual audio signal). Such a signal representation allows only a suboptimal quantization (in the sense of rate versus distortion) since a signal is generally not stationary within a block and audible artefacts such as pre-echoes or musical noise can occur [1, 3].

In our coding approach, we use a block-free signal representation and utilize a signal-adaptive decimation (i.e., subsampling) method, thus bypassing the suboptimality of block-based and constantly decimated processing. Furthermore, since our approach combines the signal representation used for the quantization with the perceptual measure, we no longer need two separate signal paths with different signal representations as common in many existing coders (e.g., the MPEG audio coders in [1]).

Finally, we note that the parameters making up the auditory representation generally are not independent. That is, coding of the auditory representation removes computational complexity associated with the distortion criterion, but it does not eliminate the need for signal modeling or other additional considerations to reduce the amount of data. In Section 4.1 and beyond, we will discuss methods that deal with this redundancy in an efficient manner.

In the next subsection, we review our auditory model, which can be inverted very efficiently to allow auditory resynthesis at high quality so that it can be used for robust coding of speech and audio signals.

1.2. An invertible auditory model

In [10] a speech coding paradigm was introduced in which the coding is performed in a perceptual domain where a simple distortion criterion (e.g., a single-letter squared error) should form an accurate and meaningful measure for the perceived distortion. In other words, the speech or audio signal is transformed into an auditory representation by passing it through an auditory model. This auditory representation is quantized and coded and the signal can be reconstructed in the decoder by an inverse auditory model.

This approach is new and different from the one used in classical perceptual audio or speech coders where an auditory model is used only in the analysis stage in parallel with the main signal path to control the quantization and bit allocation [1].

The proposed paradigm requires a model of the human auditory system that satisfies the following requirements:

(1) it provides an accurate quantitative description of perception;
(2) it leads to an auditory signal representation with relatively few parameters (to have a good basis for data compression);
(3) it can be inverted with a relatively low computational effort.

An invertible auditory model that satisfies these requirements was proposed in [10]. It is depicted in Figure 1.

In this model, the first stage is a nondecimated analysis filterbank that simulates the motion of the basilar membrane caused by acoustic stimulation. It is well known that stimuli with different frequencies produce responses with maxima at different locations along the basilar membrane. For this purpose, a functional model consists of a bank of bandpass

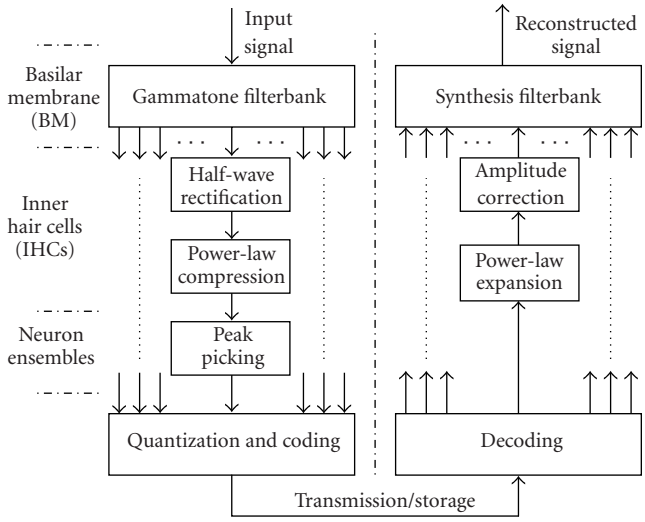

Figure 1: Invertible auditory model.

filters with different center frequencies. Note that in a human cochlea, about 2 500 inner hair cells [11] are located along the basilar membrane and, therefore, this is the actual number of bandpass channels. One reason for this high redundancy is to be robust against damages such as loss of hair cells. But this also means that neighboring auditory filters would look rather similar and, hence, for modeling purposes or coding applications, it is not necessary (and hardly possible) to implement such a high number of cochlea channels. For the invertible model in [10], the well-known gammatone filterbank with 20 channels for 8 kHz-sampled speech is used.

In each auditory channel, the analysis filterbank is followed by a model of an inner hair cell. The task of the inner hair cells is to convert the displacement of the basilar membrane in electrical receptor potentials. These receptor potentials cause a release of neurotransmitters and excite the peripheral terminals of cochlear-afferent neurons [12, 13]. In our model, this transduction process is reproduced in a very simplified way using static nonlinearities only, namely, a half-wave rectifier and a compressive nonlinearity.

The final stage in our invertible model mimics the behavior of an ensemble of cochlear-afferent neurons in each auditory channel. According to the excitation by neurotransmitters, these neurons produce action potentials ("firing pulses") caused by depolarization of an auditory nerve fiber. We model this generation of pulses using a peak-picking procedure. The set of firing-pulse trains obtained from all auditory channels is referred to as the auditory representation which is a perceptual time-frequency representation of the original speech or audio signal.

In the next section, we will describe the components of our auditory model in more detail. We cover the basilar membrane, inner hair cells, and first neural stages, that is, we model the cochlea and the auditory nerve in the human inner ear but skip the outer and the middle ear. We deal with filterbanks whose characteristics are matched to the acoustical and mechanical behavior of the cochlea and basilar membrane. One of these characteristics is that the spectral resolution decreases with increasing frequency. Therefore, warped frequency scales have been introduced long ago where selectivity bandwidths remain approximately constant along the frequency axis (auditory scales), for example, the Bark (critical-band rate) scale [14] or the ERB (equivalent rectangular bandwidth) rate scale [15]. We give a survey of auditory scales and auditory filters. The emphasis is placed on invertibility so as to allow reconstruction of the input signal. This enables the filterbank pair—analysis and synthesis filterbank—to be used for auditory subband coders or to be used in an invertible auditory model. Furthermore, we will consider important aspects for the implementation of the auditory filterbank, which is the most complex component in our model.

In Section 3 we describe the computationally efficient, nonrecursive inversion procedure of our auditory model which allows to reconstruct the input signal at a high quality from the auditory representation. We investigate our analysis/synthesis system using frame theory, which provides us with a bound for the reconstruction error.

Section 4.1 deals with the compression and quantization of the auditory representation obtained by our model and summarizes the first approaches towards low-bit-rate coding.

Since the auditory representation is highly overcomplete and does not rely on a hierarchical signal decomposition, it can be used directly for multiple description coding. We review the incorporation of our auditory model into this joint source-channel coding strategy in Section 4.2.

2. AUDITORY ANALYSIS

We selected the components of the proposed auditory model based on existing knowledge of the human auditory system. In this section, we provide additional detail for the motivation of our choices.

2.1. Basilar membrane filterbank

The filterbank to simulate the behavior of the basilar membrane is the most complex component in our model. After providing an overview of auditory filters, we consider different aspects for the implementation of an auditory filterbank.

2.1.1. Brief overview

The frequency selectivity of the human auditory system has been studied by means of psychoacoustic experiments and measurements in the cochlea and on the auditory nerve over many decades. The results of these experiments have led to the concept of auditory filters. For a historical overview, we refer to [16].

Once the bandwidths of these filters are found and expressed as a function of the center frequency, an auditory scale can be defined by integrating the reciprocal of the bandwidth function (the bandwidth function can be seen as the first derivative of the frequency with respect to the unit of the

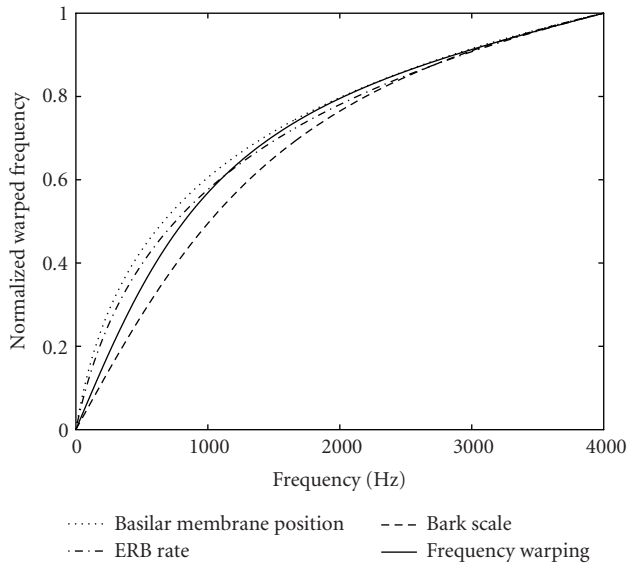

Figure 2: Comparison of the frequency-position mapping [17], the ERB rate [15], the Bark scale [50], and the frequency warping (see Appendix A.2) with $\lambda = 0.5$ for a sampling rate of 8 kHz.

bandwidth). For instance, the equivalent rectangular bandwidth $\mathrm{ERB}(f_c)$ as a function of the filter's center frequency f_c in Hz is [15]

$$\mathrm{ERB}(f_c) = 0.1079 f_c + 24.7, \tag{1}$$

and the corresponding frequency scale, the ERB rate (or "number of ERBs"), is then

$$\begin{aligned} \#\,\mathrm{ERBs}(f) &= \int \frac{df}{\mathrm{ERB}(f)} + \mathrm{const} \\ &= 21.4 \log_{10}(1 + 0.00437 f), \end{aligned} \tag{2}$$

where the integration constant has been chosen to make $\#\,\mathrm{ERBs}(0) = 0$.

Auditory frequency scales are related to the frequency-position mapping performed by the cochlea. In Figure 2, the ERB rate and the Bark [14] scales are compared with a position-frequency function which has been derived by Greenwood [17] from measurements of the mechanical motion of the basilar membrane. For more details, see [18]. In this comparison, the scales are normalized. At the maximum presented frequency of 4000 Hz, the basilar membrane position is 23.4 mm, the ERB rate reaches 27.1 ERBs, and the Bark scale has 18 Bark.

The shape of the auditory filters has been obtained by fitting different parametric expressions to experimental data. A simple linear frequency-domain description of auditory filters is the rounded exponential "roex(p, r)" function [19]

$$|H(f)|^2 = (1 - r)(1 + pg)\mathrm{e}^{-pg} + r, \tag{3}$$

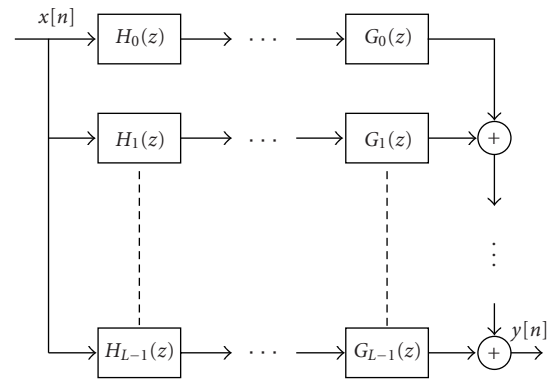

Figure 3: Analysis and synthesis filterbanks.

where g is the normalized deviation from the center frequency f_c:

$$g = \frac{|f - f_c|}{f_c}. \tag{4}$$

The parameter p determines the bandwidth and should be chosen as $p = 4 f_c / \mathrm{ERB}(f_c)$. The second parameter r flattens the shape outside the passband.

A more recent, time-domain description is the well-known gammatone function [20] for the filter impulse response

$$h(t) = t^{(l-1)} \mathrm{e}^{-2\pi b t} \cos(2\pi f_c t) \quad \text{for } t > 0, \tag{5}$$

where f_c is the frequency of the carrier and, therefore, the center frequency of the filter, b largely determines the bandwidth, and l is the order. Patterson [20] determined the choice $l = 4$ and $b = 1.019\,\mathrm{ERB}(f_c)$. For our simulations, we will use gammatone filters since the time-domain description allows straightforward FIR filter design. We will discuss this issue in more detail in the next subsection.

Several nonlinear effects have been described such as the dependency on the sound pressure level [21] which causes more asymmetric shapes of the frequency responses. For this reason, both filter descriptions have been extended [15, 22] to account for this dependency. For simplicity, particularly with respect to invertibility, we will only consider linear filters for which the above descriptions are valid for moderate sound pressure levels.

2.1.2. Implementation aspects

An implementation of an auditory filterbank consists of many auditory filters with different center frequencies in parallel. For coding applications, we should be able to reconstruct the input signal from the channel signals and the filter bank should be invertible. We denote the analysis filters as $H_k(z)$ for $k = 0, \ldots, L-1$ and the synthesis filters as $G_k(z)$ for $k = 0, \ldots, L-1$. We thus obtain the analysis-synthesis structure shown in Figure 3. Filterbank inversion and the design of synthesis filters are described in more detail in Section 3.2.

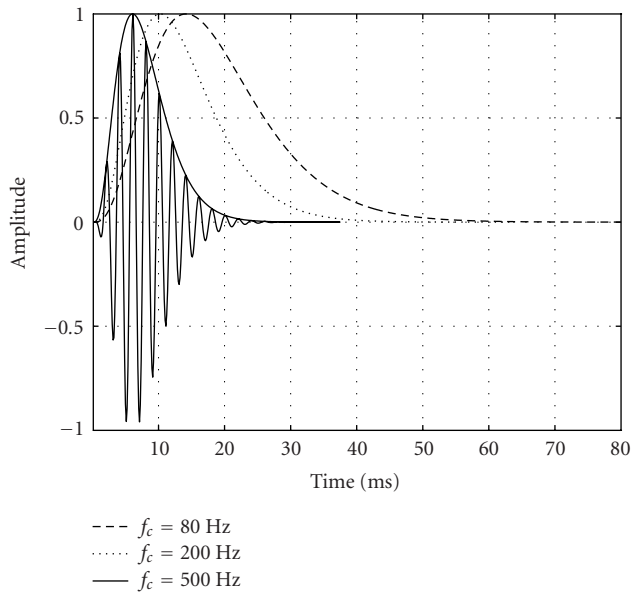

Figure 4: Impulse response and impulse response envelopes of gammatone filters for different center frequencies.

A commonly used method to compute the proper center frequencies for the filters is to transform the minimum and the maximum center frequency of interest from Hz into ERB rate. This range is divided into $L-1$ uniform sections and the obtained ERB rates are finally transformed back into Hz.

The discrete-time impulse responses of the gammatone filters can be designed by sampling and windowing the continuous-time infinite-length impulse responses of (5). A problem with direct usage of these impulse responses for FIR implementations is that the impulse responses are very long. In Figure 4, a gammatone impulse response for a center frequency of 500 Hz is plotted. Its envelope is shown as well and compared with the envelopes obtained for center frequencies of 200 and 80 Hz. As it can be seen from this figure, an impulse response with about 400 samples is needed at a sampling rate of 8 kHz for a center frequency $f_c = 200$ Hz to approximate accurately the frequency response of an ideal gammatone filter. For lower center frequencies, the length increases further (e.g., 600 samples for $f_c = 80$ Hz). Therefore, the corresponding FIR implementations are computationally expensive and memory consuming.

In the appendix, we discuss alternative implementation methods, which are computationally less expensive and should, therefore, be preferred when real-time applications running on a DSP are considered. However, for the experiments and simulations described in the following sections, we use FIR gammatone filters because computational complexity was not an issue.

2.2. Inner-hair-cell model

The auditory filterbank is followed by a half-wave rectifier and a power-law compressor, simulating the behavior of inner hair cells. The task of the inner hair cells is the so-called transduction process, that is, to convert mechanical movements into electrical potentials. It is assumed that the displacement of the cilia of the cells is proportional to the basilar membrane velocity [21]. Measurements of electrical responses have revealed a directional sensitivity: while displacement in one direction is excitatory, movement in the opposite direction is inhibitory [21]. Thus, the cells mainly react to positive deflection of the basilar membrane and, consequently, it is reasonable to model this behavior with a half-wave rectifier. Half-wave rectification is commonly used to model this aspect of physiology, for example, [4, 23, 24].

The aforementioned measurements also show a compressive response [21]. Therefore, we apply a power-law compressor to the half-wave rectified signals. The input $x[n]$ and the output $y[n]$ of the compression stage are related by

$$y[n] = x^c[n], \tag{6}$$

with $c = 0.4$. This stage is similar to logarithmic amplitude compression schemes in ordinary waveform coders (e.g., μ-law).

The static nonlinearity is a strongly simplified model of the human peripheral processing. In related literature, more sophisticated compression or adaptation stages have been proposed. In [4], a cascade of five feedback loops with different time constants is used. The cascade compresses stationary sounds almost logarithmically whereas rapidly varying signals are transformed more linearly, thus modeling the "overshoot effect," that is, a higher sensitivity at the onset of a stimulus. Other examples can be found in [23, 24] where automatic gain controllers model the synaptic region between the hair cell and the nerve fiber. In our first implementation of an invertible model, we use the simple half-wave rectifier and power-law compressor to avoid stability problems when inverting the gain control loops.

2.3. Neuron model

Contrary to many other auditory models (e.g., [4, 23, 24]), we preserve the temporal fine structure of the signal, that is, we do not apply time averaging to the subband signals because this would lead to a low reconstruction quality. In our model the power-law compressor is followed by an adaptive subsampling mechanism ("peak picking"), which searches for local maxima and sets all other samples to zero. Let the input and the output of the peak-picking stage be denoted by $x[n]$ and $y[n]$, respectively, then the output can be calculated as

$$y[n] = \begin{cases} x[n], & x[n] > x[n-1] \wedge x[n] > x[n+1], \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

This model simulates the firing behavior of an ensemble of auditory neurons. The responses are clusters of high firing activity that are synchronized (phase-locked) with the waveform shape of the input signal.

It is known that a single neuron generally does not fire more often than 250 times per second [12, 13] and, therefore, it is by itself not able to preserve the time structure of

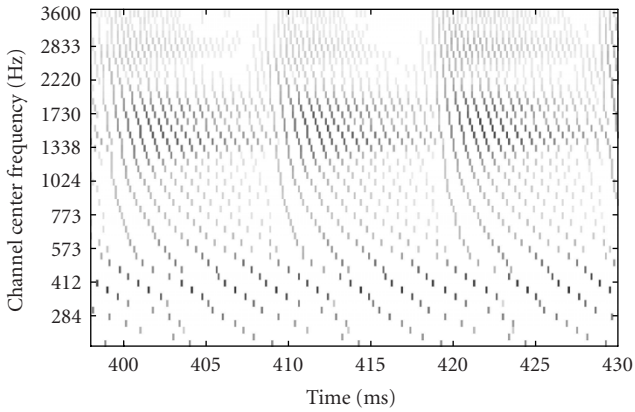

Figure 5: Auditory representation (here with 50 channels) of the sound [I] taken from "there is," spoken by a male. Peaks are shown as rectangles with their intensity representing their amplitude. The time axis covers three pitch periods.

high-frequency components. Since in the early human auditory system, about 30 000 neurons [11] encode the signals of significantly less hair cells, we can associate several neurons with one hair cell output. Our model of the neurons is physiologically plausible. Each neuron has an internal state that decays exponentially with a relatively large time constant. When it fires, this state is reset to a value that depends on the input signal level. The firing probability increases monotonically with the difference between the neuron's input and its state. So an ensemble of neurons shows a high firing rate at the peak of the input signal. The amplitude of a pulse in our model represents the firing rate, that is, the number of neurons of the ensemble that fire at the peak location.

The effect of phase locking is known to occur only at frequencies below 4 kHz [12, 13]. So the used model is physiologically plausible for the coding of narrowband speech signals. For simplicity, we use this neuron model even if we process signals at higher sampling rate, for example, wideband speech or general audio signals.

The consideration of pulsed neural models where information is carried in the pulse timings is clearly motivated by observations of biological neural networks. In [25] it is well demonstrated that these models should be preferred to classical neuron models such as firing-rate models or even more simplified ones for many applications of artificial neural networks.

In Figure 5, a pulse representation of a segment of about 30 milliseconds duration taken from a voiced speech is shown. For this example, a 50-channel FIR gammatone analysis filter bank was used. The neuron firings are not strictly aligned across the frequency channels due to different delays of the filters. Nevertheless, the phase-locking effect can be seen clearly. Also the formant structure is visible with formants around 400 Hz, 1700 Hz, and 2800 Hz.

Weintraub [26] used a similar deterministic model for neural firing in his sound separation system. There is also similarity to Patterson's pulse ribbon model [27] but we preserve the amplitudes of the pulses in addition to the locations. Contrary to [26, 27], we are able to resynthesize the original audio signal directly from this neural firing pulses whereas Weintraub uses the (unprocessed) signals from the auditory filterbank for the resynthesis [28] and Patterson does not resynthesize at all.

3. AUDITORY SYNTHESIS

The attempts of resynthesis of the input signal from an auditory representation are not new. In [29] a historical overview is given. The aim of various model inversions was to understand perception [30, 31, 32], to test the accuracy of the model [33, 34], and to separate speech from noisy backgrounds or interfering speakers [26, 28, 32]. We propose to use an invertible auditory model for coding of speech and audio signals [10, 35].

For the most recent models, the inversion method is based on projections onto convex sets [32, 34] and utilizes iterative signal reconstruction algorithms. The resynthesis of our auditory model does not need iterative procedures and is, therefore, computationally very efficient and nevertheless perceptually accurate.

3.1. Inversion of neuron and inner-hair-cell models

The first step in the inversion procedure is to undo the power-law compression using the proper inverse expansion to get the positive peak amplitudes of the original subband signal:

$$y[n] = x^{1/c}[n]. \tag{8}$$

Now, each of the channel signals approximates the situation where a signal is downsampled and then upsampled by means of inserting zeros. This insertion of zeros leads to aliasing which can be removed by bandpass filtering. The bandpass filters are located in the synthesis filterbank. Before they are applied, the amplitude of the pulses has to be corrected to compensate for the loss of energy due to (i) the adaptive downsampling and (ii) the peak amplitude errors at higher frequencies introduced by the finite sampling rate.

We consider one auditory channel. The output of one channel of the analysis filterbank resembles a sinusoid with a period of P samples that is related (but not identical) to the inverse of the center frequency of the filter. Then the peak-picking procedure behaves like a cascade of an ordinary downsampler and upsampler with a fixed decimation/interpolation factor P for which the Fourier transform relation is

$$Y(\mathrm{e}^{j\theta}) = \frac{1}{P}\sum_{k=0}^{P-1} X(\mathrm{e}^{j(\theta-k2\pi/P)}). \tag{9}$$

The cosine signal with amplitude 1 and angular frequency $2\pi/P$ with Fourier transform

$$X(\mathrm{e}^{j\theta}) = \pi\left(\delta_{2\pi}\left(\theta - \frac{2\pi}{P}\right) + \delta_{2\pi}\left(\theta + \frac{2\pi}{P}\right)\right) \tag{10}$$

is transformed into the pulse train with Fourier transform

$$Y(\mathrm{e}^{j\theta}) = \frac{2\pi}{P}\sum_{k=0}^{P-1} \delta_{2\pi}\left(\theta - \frac{k2\pi}{P}\right), \tag{11}$$

where $\delta_{2\pi}(\theta)$ is the 2π-periodic delta distribution. All additional frequency components have to be attenuated by the synthesis filter and the remaining components yield the cosine signal with amplitude $2/P$. Therefore, the amplitude in this channel has to be corrected by a factor of $P/2$. This method is very simple and contributes substantially to good resynthesis results. Another slightly more elaborate correction method is to count the actual number of zeros between adjacent pulses which replaces the constant correction factor with an adaptive one.

For the second correction step, we observe that the measurement of the peak amplitude is exact in continuous time only. In discrete time, errors due to the finite sampling interval are inevitable. These errors become significant in particular for those auditory channels whose center frequencies are close to half the sampling frequency. To compensate for these errors, a method based on the assumption of a uniformly distributed random sampling error was proposed in [10]. The method evaluates the average per-cycle maximum amplitude of a sampled sinusoid, α, which, for the case of a unity amplitude sine wave and a unity sampling period, is given by

$$\alpha = \int_{-1/2}^{1/2} \cos\left(\frac{2\pi t}{P}\right) \mathrm{d}t = \frac{P}{\pi} \sin\left(\frac{\pi}{P}\right). \tag{12}$$

Thus, the correction factor due to the finite sampling rate for this channel is $1/\alpha$.

An improved correction factor was introduced in [35] which is based on least-squares optimization. For a sinusoidal signal with amplitude A and period P, we observe the maximum sample $w_{\max} = A\cos(2\pi t/P)$ with t uniform over $[-1/2, 1/2]$. The nonlinear least-squares estimate for the amplitude $\widehat{A}$ in terms of the observation $w_{\max}$ is given by $\widehat{A} = \mathrm{E}\{A|w_{\max}\} = \beta \cdot w_{\max}$ with

$$\beta = \int_{-1/2}^{1/2} \frac{1}{\cos(2\pi t/P)} \mathrm{d}t = \frac{P}{\pi} \ln\left[\tan\left(\frac{\pi}{4} + \frac{\pi}{2P}\right)\right]. \tag{13}$$

In Figure 6, these two compensation methods are compared. For a white-noise input signal, the power spectral density function of the output signal is plotted for the cases of no peak picking and therefore no correction ("nondecimated case"), peak picking with correction by $1/\alpha$, with correction by β, and peak picking without correction. We recognize that the correction factor β according to (13) keeps the reconstruction error less than 1 dB across the entire frequency range covered by the auditory filterbank.

3.2. Synthesis filterbank

The last stage is the synthesis filter bank, which should be an inverse of the analysis filterbank. For proper signal reconstruction from a firing-pulse representation, it is essential that the synthesis filters have bandpass characteristics to eliminate aliasing. This also keeps the introduced quantization noise within a local frequency range.

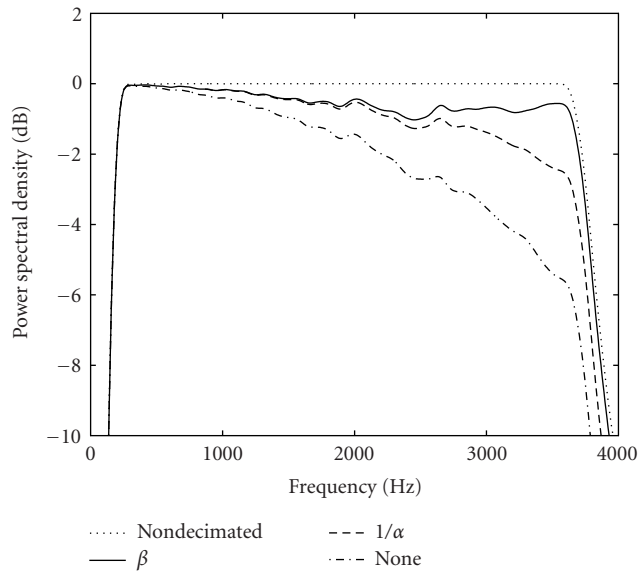

Figure 6: Comparison of reconstruction quality with different amplitude correction methods for the peak amplitude sampling errors (output PSD for white input).

In general, the inverse operator for a nondecimated, invertible filterbank is not unique. A natural method of inversion of a nondecimated FIR filterbank is based on the following condition for perfect reconstruction:[1]

$$G_k(z) = \frac{H_k(z^{-1})}{\sum_{i=0}^{L-1} H_i(z) H_i(z^{-1})}. \tag{14}$$

For the case that $\sum_{i=0}^{L-1} H_i(z)H_i(z^{-1}) = 1$, the synthesis filterbank is the analysis filterbank with time-reversed impulse responses. A delay equal to the length of the analysis filters minus one is needed to make the synthesis filterbank causal.

In the general case, when the denominator of (14) is not equal to one (e.g., when a low number of auditory channels is used), accurate signal reconstruction can be obtained with an additional linear-phase equalization filter (see [10]) that operates on the sum of all channels synthesized with $G_k(z) = H_k(z^{-1})$. This equalizer has to be designed to approximate the frequency response

$$E(\mathrm{e}^{j\theta}) \stackrel{!}{=} \left[\sum_k \left|H_k(\mathrm{e}^{j\theta})\right|^2\right]^{-1} \tag{15}$$

to reduce the remaining magnitude ripple.

The ripple decreases with increasing order of the FIR equalizer. However, an additional delay of half the filter order is introduced. Thus, for the choice of the impulse response length, a suitable compromise must be found. The minimum

[1] Here, perfect reconstruction refers to processing of the input signal by the analysis and the synthesis filterbank only.

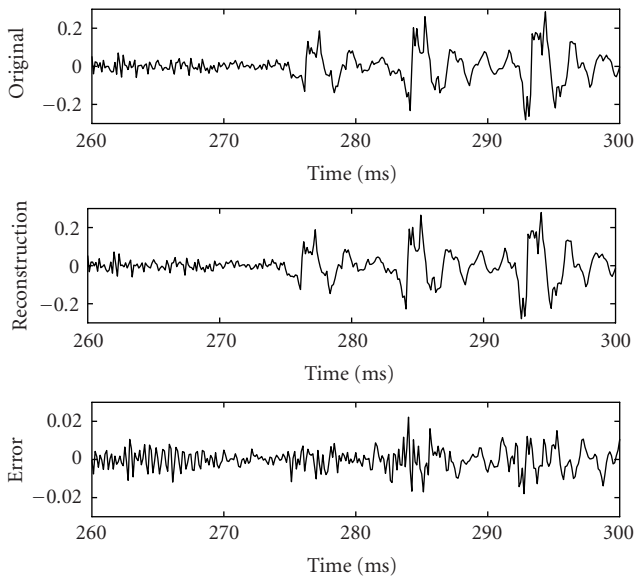

Figure 7: Comparison of the original waveform (upper plot), the waveform reconstructed from the auditory representation (middle plot), and the reconstruction error (lower plot, note the finer amplitude scale). Speech segment taken from "The *so*urce," spoken by a male speaker.

delay solution without equalization has often been used [32]. We found that, for 20 channels and for a sampling rate of 8 kHz, this results in a 4 dB ripple. The ripple decreases with a further increase of the number of channels.

As already mentioned for the analysis filters, FIR gammatone filters are memory consuming. Although the synthesis filters can use the same coefficients as used for the analysis filters, separate ring buffers are needed for every auditory channel in the synthesis filterbank. Consequently, the necessary amount of memory is doubled. For an accurate FIR gammatone filterbank implementation with long impulse responses, the memory of the most currently used DSPs is not sufficient. One solution to this problem is to take shorter impulse responses and accept deviations from the ideal frequency responses. Another possibility is to consider alternative filterbank implementations as described in the appendix.

3.3. Simulation results

In Figure 7, a segment of the original waveform with 8 kHz sampling rate is compared with the output of our inverse auditory model with 20 channels. For this simulation, the FIR gammatone filterbank from Irino's Matlab toolbox[2] was used where the lowest center frequency was 100 Hz (order of filter 666) and the highest 3600 Hz (order 56). The auditory representation has been left uncompressed. The output of the synthesis filterbank has been passed through a linear-phase equalizer with a group delay of 25 milliseconds. Although the average segmental signal-to-noise ratio (SNR) is only 17.9 dB, the reconstructed signal is without audible distortion (evaluated by two experienced listeners in an original/reconstructed-comparison listening test).

3.4. Frame-theoretic interpretation of auditory synthesis

It is useful to consider the auditory resynthesis from the perspective of frame theory. This endorses our choice of synthesis filterbank and provides a bound for the reconstruction error introduced by the analysis/synthesis filterbank pair. Furthermore, it justifies our simple method to reconstruct the signal from the pulse representation and allows us to reduce the number of pulses in the auditory representation.

In practical implementations of the filterbank structure, the analysis and synthesis filterbanks are identical, except for a time reversal of the impulse responses. We first evaluate the validity and implications of this choice. The analysis filterbank maps the input sequence[3] to a set of channel sequences, one for each filter. It is essential that the analysis filterbank is invertible and that means it can be interpreted as a frame operator, which we denote as F. The analysis filterbank operation can be written as a set of inner products, denoted as $(Fx)[j] = \sum_i \psi_j^*[i]x[i]$, with functions $\{\psi_j\}_{j\in J}$ where each is a translate of one of the L time-reversed impulse responses. The indexes j enumerate each output sample of all L channels. Invertibility of the filterbank is guaranteed if the frame condition is satisfied:

$$A\sum_{i\in\mathbb{Z}}\left|x[i]\right|^2 \le \sum_{j\in J}\left|(Fx)[j]\right|^2 \le B\sum_{i\in\mathbb{Z}}\left|x[i]\right|^2 \quad \forall x\in\ell^2(\mathbb{Z}), \tag{16}$$

where A and B are finite, positive, scalar frame bounds. The adjoint operator F^* maps an L-channel signal, y, to a single-channel signal $F^*y = \sum_{j\in\mathcal{J}} y_j\psi_j$.

In general, the inverse frame operator (the synthesis filterbank) is not unique. We are interested in an inverse that is easy to compute, and, importantly, that minimizes the effect of quantization errors in $(Fx)[j]$ on the reconstruction. The so-called frame algorithm is an iterative procedure that provides the inverse frame operator that minimizes the effect of quantization errors. The first iteration often provides a useful approximation to the inverse or even the exact inverse. The estimate x_m of x at iteration m of the frame algorithm is

$$x_m = \rho F^*y + (\mathrm{Id} - \rho F^*F)x_{m-1}, \tag{17}$$

where ρ is a scalar relaxation parameter, Id is the identity operator, and $x_0 = 0$. The estimation error at iteration m is then

$$x - x_m = (\mathrm{Id} - \rho F^*F)(x - x_{m-1}) = (\mathrm{Id} - \rho F^*F)^m x. \tag{18}$$

[2]This toolbox can be found at http://www.mrc-cbu.cam.ac.uk/cnbh/aimmanual/.

[3]We assume that the input sequence is in the Hilbert space $\ell^2(\mathbb{Z})$.

With the optimal selection $\rho = 2/(B+A)$, the error is bounded by

$$\begin{aligned} ||x - x_m|| &= \min_{\rho} ||(\mathrm{Id} - \rho F^* F)^m x|| \\ &\leq \min_{\rho} \max\left(|1-\rho A|, |1-\rho B|\right)^m \|x\| \\ &= \left(\frac{B-A}{B+A}\right)^m \|x\|. \end{aligned} \tag{19}$$

The values A and B form the minimum and maximum eigenvalues of the operator F^*F, which are precisely the frame bounds.

The first-iteration estimate of x by the frame algorithm is the expansion $\rho F^* y = \rho \sum_{j\in\mathcal{J}} y_j \psi_j$, which implies that ρF^* is the approximation to the inverse operator. It is easily seen that this corresponds to a synthesis filterbank with impulse responses that are the time-reversed impulse responses of the analysis filterbank, scaled by ρ. Moreover, we see from (19) that the relative error is bound by the factor $(B-A)/(B+A)$.[4]

For a nondecimated filterbank, the discrete-time Fourier transform (which is unitary) simplifies the analysis of the operator F^*F. In the Fourier domain, the operator F^*F corresponds to the operator [36]

$$\mathcal{F} F^* F \mathcal{F}^{-1} = \sum_{i=0}^{L-1} H_i(\mathrm{e}^{j\theta}) H_i(\mathrm{e}^{-j\theta}), \tag{20}$$

where $\mathcal{F}$ denotes the discrete-time Fourier transform operator. This immediately leads to the inversion formula given in (14). The same Fourier-domain equivalence shows that the frame bounds then correspond to the essential infimum and supremum of $\sum_{i=0}^{L-1} H_i(\mathrm{e}^{j\theta}) H_i(\mathrm{e}^{-j\theta})$.

We can now draw some conclusions for our auditory filterbanks based on the frame-theoretical viewpoint. First, the synthesis filterbank based on time-reversing the impulse responses is an approximation to the perfect synthesis filterbank that has minimum sensitivity to quantization errors in the perceptual domain. Second, the accuracy of this approximation is governed by the relative error $(B-A)/(B+A)$, where A and B can be evaluated as the essential infimum and supremum of the summed responses of the analysis filterbank. For an auditory filterbank implementation based on FIR gammatone filters, the relative error $(B-A)/(B+A)$ is -30.7 dB for 50 channels and -5.9 dB for 20 channels.

Frame theory can also be used to provide an interpretation of the peak-picking procedure that we use in our auditory model. It is convenient to look at a single channel first. A frame algorithm that can be used for the reconstruction of continuous lowpass band-limited signals from irregularly spaced samples and their derivatives was presented in [37]. In this case, the frame is formed by the translates of the impulse response of an ideal lowpass filter and its derivatives. For our case, the first-order derivative of the signal samples is selected as zero and the reconstruction method is essentially identical to the reconstruction applicable if no derivative is given. However, reconstruction is possible with a larger spacing between the samples than if no information was known about the derivatives (a factor two for regularly spaced samples). In practice, the first iteration of the frame algorithm consists of ideal lowpass filtering of the upsampled (inserting zeros) weighted signal. The weighting of each sample is linear with the distance to the previous sample. Nearly uniform spacing, as we have in our case, results in nearly uniform weighting, reducing the first iteration of the frame algorithm essentially to a lowpass filter. Moreover, it is easy to see that the frame is tight for the regular sampling case, which means that the first iteration renders the exact inverse.

We note that the frame algorithm of [37] assumes a band limited signal and a sample spacing that is at most $2\pi/\theta$ for a band limitation of θ (in practice, the band limitation is somewhat less). Since the output of the auditory filters resembles sinusoids, and since a sinusoid of frequency θ_s has its maxima spaced at $2\pi/\theta_s$, this implies that the frame algorithm of [37] does not apply to our case without modification. The required modification consists of replacing the impulse response of the ideal lowpass filter by the impulse response of an ideal bandpass filter.[5] For regularly spaced samples, the reconstruction algorithm then consists of a simple bandpass filtering. For irregular spacing, the samples must first be weighted appropriately.

In practice, the bandpass filtering operation required for the reconstruction of each of the irregularly sampled channels can be usurped by the corresponding synthesis filter within the inverse of the basilar membrane filterbank. In our practical implementation, we then make the following approximations with respect to inverting the peak-picking procedure: (i) we use the first iteration of the frame algorithm and this is not accurate since the frame is not tight for irregular sampling, (ii) we neglect the sample weightings that are needed to account for irregular sampling, and (iii) we assume the narrowband character of the inverse basilar membrane filterbank filters allow the bandpass filters to be omitted. The perceptual effect of these approximations on auditory synthesis is small; the samples are almost uniformly spaced and the bandpass filters used to invert the peak picking can be very broad, broader than the auditory filters, as is confirmed by the results provided in Section 3.3.

The frame interpretation leads directly to a method to reduce the coding rate of our basic model. Particularly for the filters of the basilar membrane filterbank with high center frequency, the peak-picking procedure leads to a high rate of peaks. Since the peak locations and amplitudes must be encoded as side information, the resulting parameterization is not a good basis for coding. However, we note that the described frame-algorithm-based reconstruction from peak

[4] The first-iteration estimate is exact for $A = B$, which corresponds to a tight frame.

[5] We note that, in general, sampling rates that are sufficient for lowpass signals may not be so for bandpass signals of identical bandwidth, for example, see [38]. However, this aliasing problem is unlikely to occur for spectra that essentially consist of a single line.

amplitudes and locations only requires that the peaks be not separated by more than a given distance. Importantly, there is no requirement to include all peaks of the signal. As a result, we can downsample the peak sequence in the channels with higher center frequency by a significant factor without loosing the ability to reconstruct the signal.

The amount of downsampling that can be applied to a peak sequence is constrained by the bandwidth of the ideal bandpass filter of the frame. With increasing downsampling of the peak sequence, the importance of the bandpass filtering operation increases and then it cannot be omitted from the synthesis structure. On the other hand, the bandpass filter cannot be selected to be narrower than the nominal width of the basilar membrane filters, since that removes relevant information. It is interesting to note that this frame-theoretical vantage point leads to a new interpretation of the results obtained in [39]. In [39], downsampling of the peak sequence was justified from a masking argument, which is not physiologically plausible for the auditory representation.

4. EXEMPLARY APPLICATIONS IN AUDIO AND SPEECH CODING

The proposed invertible auditory model allows to resynthesize the input signal with high quality and, therefore, can build a basis for coding of audio signals. The next section describes first approaches for quantization and coding to reduce the amount of data needed to transmit an auditory representation, whereas in Section 4.2, we exploit the inherent redundancy of the auditory representation in a joint source-channel coding strategy to protect against possible losses of data during the transmission in a packet-switched network.

4.1. *Auditory-domain compression*

The auditory representation provided by our model is sparse, consisting mostly of zeros. However, it contains more firing pulses in total compared to the number of samples that the original input signal has (about three times more for the 20-channel case and a sampling rate of 8 kHz).

Experiments have shown that the firing-pulse amplitudes can be quantized coarsely, for example, using a block scalar quantizer with a block duration of 20 milliseconds and 1 bit [10] per pulse amplitude, without introducing audible distortion. The maximum amplitude of a block has to be transmitted as side information for each channel where 6 bits per value are enough. In fact, quantizing the peak amplitudes with 1 bit enables us to refer to three amplitude values—high (= 1), middle (= 0), and zero—since zero denotes that there is no pulse at all (no pulse time position transmitted as side information).

In [39] even 0 bits were found to be sufficient for the pulse amplitudes, that is, only the side information (block-average pulse amplitudes and pulse positions) requires transmission. However, in that work shorter block lengths are used to determine the block energy. Especially for higher-frequency channels, the block duration is only 4 milliseconds.

Much more important than the firing-pulse amplitudes are the pulse-time positions. Also these positions have to be transmitted as side information which produces by far the major part of the transmitted data. In [39], these positions are compressed using arithmetic coding for low-frequency channels and vector quantization for high-frequency channels. Furthermore, models of temporal and simultaneous masking were added to reduce the overall number of firing pulses drastically. While the consideration of simultaneous masking does not bring a remarkable reduction, exploiting temporal masking does. Our own experiments with the model for temporal postmasking adopted from [39] show that an average reduction in number of pulses by 50% for 16 kHz-sampled speech does not affect the audible quality of the reconstructed signal [40]. For this model, a masking threshold signal is computed in each channel. Let $x[n]$ be the firing pulse train of one auditory channel and $T[n]$ the corresponding masking threshold. Then $T[n]$ is defined as

$$T[n] = \begin{cases} x[n], & x[n] > T[n-1]\mathrm{e}^{-1/\tau}, \\ T[n-1]\mathrm{e}^{-1/\tau}, & \text{otherwise.} \end{cases} \tag{21}$$

The time constant τ was set to 125 samples for the lowest-frequency channel and to 33 samples for the highest (according to the empirically determined values from [39]). Once this threshold is computed, the output signal of the masking stage is

$$y[n] = \begin{cases} x[n], & x[n] > T[n-1]\mathrm{e}^{-1/\tau}, \\ 0, & \text{otherwise.} \end{cases} \tag{22}$$

It is natural to observe many more pulses in high-frequency channels[6] than in low-frequency channels. Thus, the reduction of the number of pulses is most effective in high-frequency channels. This is in accordance with the frame-theoretic consideration of Section 3.4.

We have performed experiments with 16 kHz-sampled speech and a 16-channel auditory model. The aforementioned temporal masking model has been included to reduce the number of pulses. The positions of the remaining firing pulses have been coded using run-length encoding combined with arithmetic coding, which results in an average bit rate of about 100 kbps [40] for the transmission of the pulse positions only.[7] Further compression can be achieved with vector quantization, (cf. [39]), where an average bit rate of about 70 kbps has been achieved for the overall bit stream.

We expect that a coarse quantization of the pulse positions in high-frequency channels should be sufficient since neurons of the auditory nerve do not any longer show phase-locked firing behavior above 4 kHz. Thus, we expect only a minor increase in necessary bit rate when audio signals at

[6] The average number of pulses per second in an auditory channel can be predicted by the channel's center frequency.

[7] The amplitude information must be added.

higher sampling rates (e.g., 44.1 kHz) are coded. To affirm this is a matter of our current research.

We have to further reduce the number of pulses significantly, particularly in higher-frequency channels, to achieve a better compression. In our most recent work [41], we incorporated a combined model for both simultaneous and temporal masking. Together with another pulse-amplitude correction step, which compensates for the loss of energy due to the elimination of pulses, we are able to omit even 74% of the original pulses of speech signals sampled at both 8 kHz and 16 kHz without degrading the reconstruction quality. This result is a step further towards an efficient compression method since it reduces the amount of side information considerably. To find the upper bound of the downsampling factor without loosing the ability to reconstruct the signal is a matter of further investigations.

4.2. Multiple description coding

The high degree of redundancy in the human peripheral auditory system forms a motivation to use our invertible auditory model in a joint source-channel coding strategy. In other words, we use the overcomplete auditory representation to protect the transmitted signal against erasure of coded information (packet loss in packet networks). In [35], we proposed the first instance of a highly redundant speech coder optimal for packet-switched networks, for example, for voice over IP applications. There, we use the auditory model for multiple-description coding where the source information is spread over multiple signal descriptions which are carried over M independent subchannels. These transport channels may be physically distinct as in a packet-switched network or correspond to multiplexed subchannels on a single physical channel. When an arbitrary set of $K < M$ subchannels fails, the receiver uses the information from the remaining $M - K$ intact channels to reconstruct the transmitted signal. Therefore, the encoder should be based on a nonhierarchical signal decomposition [42]—this is the case for our auditory representation—and should assign descriptions of equal importance to each transport channel. The descriptions must be different, that is, each must carry new information, such that receiving more descriptions enables the decoder to improve the reconstruction quality.

A grouping of the L auditory channels into $M \leq L$ transport channels provides an immediate application of our coding paradigm in this context. To form descriptions of equal importance, L should be chosen as an integer multiple of M such that a constant number of L/M uniformly spread auditory channels are packaged together into one transport channel. In this respect, each description is obtained by frequency-domain subsampling of an overcomplete signal representation. One extreme case is given if $M = L$, that is, the maximum possible number of transport channels is used to achieve superior robustness. The other extreme case is a simple interleaving of odd and even indexed auditory channels and assigning them to $M = 2$ transport channels.

If erasures occur, the coded information about some auditory channels is lost. However, the information at the affected frequencies is generally not lost because neighboring auditory filters overlap. Assuming that the decoder knows which channels are erased, as is the case in packet networks, a time-varying equalizer filter can be designed for the reconstruction after the synthesis filterbank to amplify the attenuated regions.

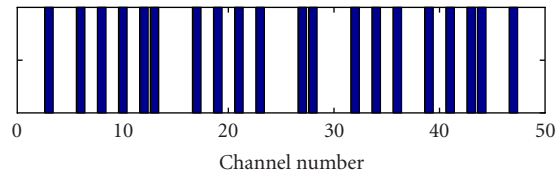

Figure 8: Channel erasure pattern for the 50-channel auditory representation. Black bars indicate erased channels (40%); white bars stand for intact ones (60%).

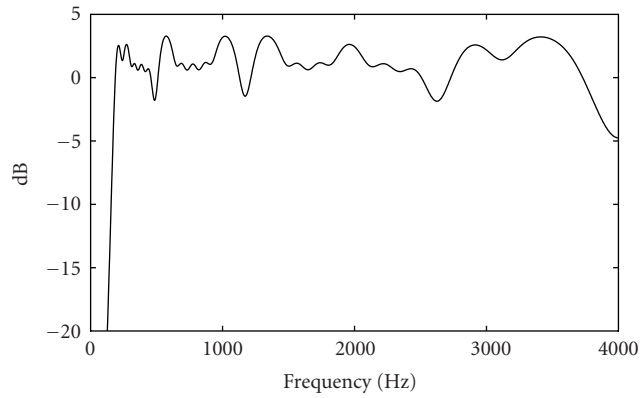

Figure 9: Overall unequalized frequency response of the nondecimated analysis/synthesis filterbank with channel erasures as in Figure 8.

From a frame-theoretical viewpoint, the perfect inverse filter bank can be constructed as long as the frame functions corresponding to the received information form a frame, that is, if they satisfy the frame condition displayed in (16). However, since the separation between the essential infimum and supremum will increase, the approximation made by using time-reversed impulse responses will become less accurate. The accuracy of the approximation prior to equalization can be quantified by means of the factor $(B - A)/(B + A)$, which bounds the distortion.

Experiment and Results

We have run an experiment with an auditory model with 50 auditory channels assigned to 50 transport channels. We generated the channel-erasure pattern with 40% randomly muted channels shown in Figure 8. Although the 50 auditory channels highly overlap, the high proportion of erased channels creates a clearly perceptible spectral distortion (cf. Figure 9) if no equalizer is used. The factor $(B - A)/(B + A)$, which bounds distortion, is -3.0 dB.

The amplitudes of the firing pulses are represented with 1 bit each using block-adaptive quantization, whereas the

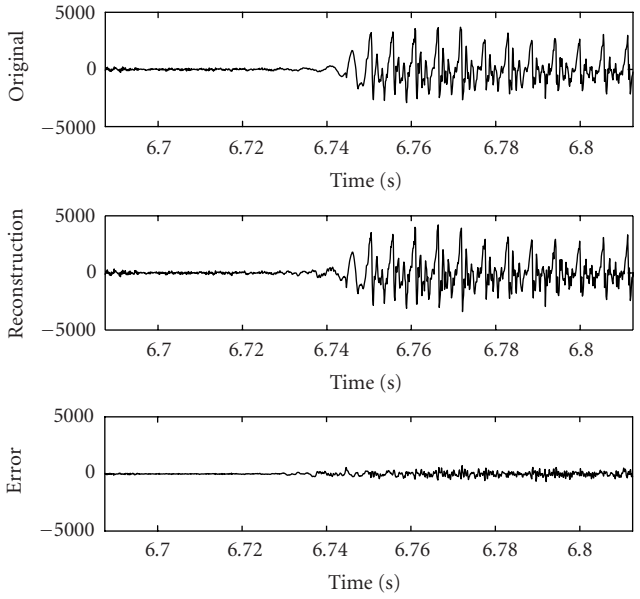

Figure 10: Original segment of a speech waveform (first plot) versus reconstruction from decimated and quantized auditory representation with 40% channel erasures (middle) and the difference between both (third plot).

pulse positions are left unquantized. In Figure 10, the reconstruction results are shown with a waveform of the initial part of the word "player" spoken by a female speaker sampled at a rate of 8 kHz. In the first plot, the original waveform (compensated for the processing delay) is drawn. The second plot shows the output of the decoder for the case that 40% of the channels are erased and an appropriate equalizer with an impulse response of length 256 samples is used. In the third plot, the reconstruction error, that is, the difference between the original and the reconstructed signal, is plotted. The average segmental SNR is 15.5 dB compared to 16.4 dB in the case without channel erasures.

These results show potential applicability of our invertible auditory model in joint source-channel coding methods such as multiple description coding for robust transmission over packet-switched networks.

5. CONCLUSION

We have reviewed an invertible auditory model and its usage for robust coding of speech and audio signals. The inversion procedure to reconstruct the original signal from its auditory representation does not need computationally expensive iterative algorithms and produces reconstructed audio signals with very high quality. The overcomplete auditory representation suggests the application of the invertible auditory model in multiple description coding. Our experiments have shown that our auditory model provides an ideal basis for this joint source-channel coding method to allow robust transmission over packet-switched networks even if a high number of packets get lost. Experiments have shown promising results.

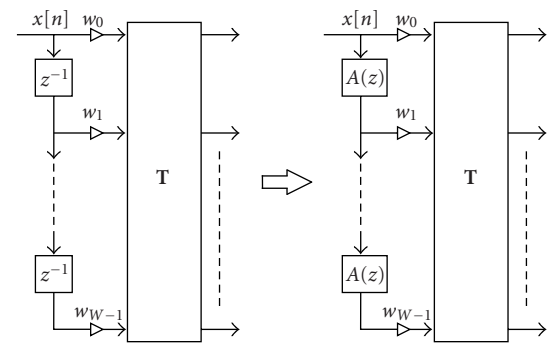

Figure 11: Modification of a nondecimated transform filterbank to obtain a frequency-warped version.

APPENDIX

A. ALTERNATIVE FILTERBANK IMPLEMENTATION METHODS

As discussed in Section 2.1.2, FIR implementations are computationally expensive and memory consuming.

A.1. IIR filterbank

Several computationally less expensive IIR implementations for gammatone filters [43] have been suggested. These are based on usual transforms from continuous-time transfer functions to discrete-time transfer functions (e.g., impulse-invariance transformation) which result in filters with an order of 8.

Inversion

An inversion based on FIR filters according to (14) is not possible for infinite impulse response filters. While for nondecimated filterbanks the direct channel-by-channel inversion of minimum-phase analysis filters is possible with stable and causal synthesis filters, this is not advisable since the frequency response of the inverse is complementary, that is, the inverse of a bandpass filter gives a bandstop. In this paper, we do not deal with further inversion possibilities for IIR filter banks, but refer to [44].

A.2. Frequency-warped transform filterbank

Another computationally very efficient approximation of an auditory filterbank is to take a frequency-warped transform filterbank. In the early 1970s, Oppenheim, etal. [45] introduced the technique of computing nonuniform resolution Fourier transforms. They first transform the input sequence into a frequency-warped version by time-reversing and passing it through a chain of allpass filters. After that, an FFT of the samples along this allpass cascade is performed. This is a computationally very efficient method for a constant relative-bandwidth spectral analysis for finite-length signals.

In the late 1970s, Vary [46] suggested a frequency-warped transform filterbank obtained by simply replacing the unit-delay elements in the signal flow graph representation of a sliding window with general allpasses. This is illustrated in Figure 11. The window coefficients $w_0, \ldots, w_{W-1}$ correspond to the impulse response of the prototype lowpass filter which

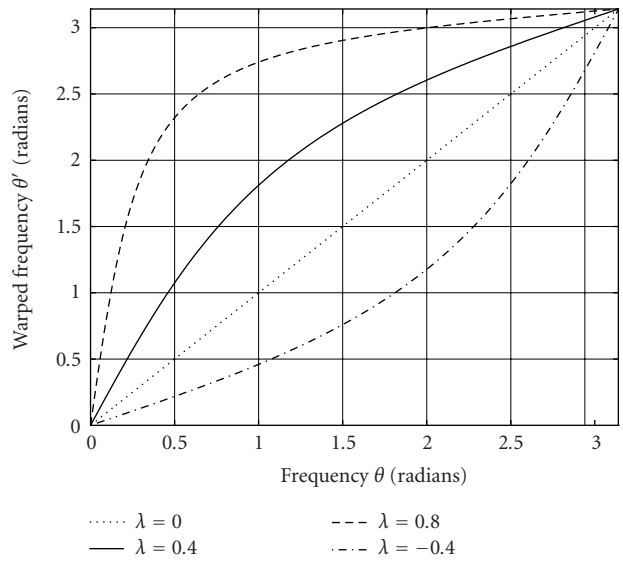

Figure 12: Phase function of the first-order allpass for four different values of the warping parameter λ.

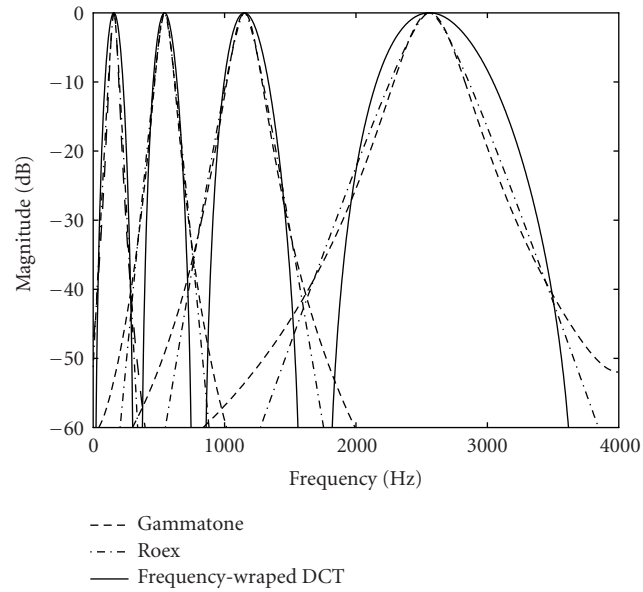

Figure 13: Normalized frequency responses of four channels of auditory filterbanks. Comparison between FIR gammatone filters, rounded exponentials, and a frequency-warped DCT-4 filterbank ($\lambda = 0.5$, $f_s = 8$ kHz, 64-point Kaiser window with $\beta = 10$).

is modulated by the transform **T** (e.g., a DFT or a DCT) to get bandpasses. The window length W does not necessarily have to be equal to the number of channels L (see [47] for more details). Thus, a longer FIR prototype filter can be designed to better approximate gammatone or roex frequency responses.

We consider a nondecimated filterbank where the window advances by one sample at a time and, thus, the transform has to be calculated for every sample and nondecimated subband signals are obtained.

When the unit delays are replaced with general nonlinear-phase allpasses, the characteristics of the transform filterbank will be modified. Let the transfer function of a first-order allpass be denoted by

$$A(z) = \frac{z^{-1} - \lambda}{1 - \lambda z^{-1}}. \tag{A.1}$$

with the single so-called warping parameter λ. If we substitute z^{-1} by $A(z)$, a bilinear transform is applied resulting in warping the frequency axis corresponding to the phase function of the allpass

$$\theta' = \arctan\left(\frac{(1-\lambda^2)\sin(\theta)}{(1+\lambda^2)\cos(\theta) - 2\lambda}\right), \tag{A.2}$$

where θ and θ' are the frequency (in radians relative to the sampling frequency) variables before and after warping, respectively. In Figure 12, this function is plotted for different warping parameters λ.

Smith and Abel [48] proposed expressions for choosing a proper λ to achieve a frequency warping nearly identical to that of the Bark or the ERB rate frequency scales for a given sampling frequency. In Figure 2, the warped frequency scale obtained using allpasses with $\lambda = 0.5$ at a sampling rate of 8 kHz is compared with the frequency-position function, the ERB rate scale, and the Bark (critical-band rate) scale. Therefore, warping a uniform filterbank with a chain of first-order allpasses yields a good approximation of auditory filterbanks for critical-band spectral analysis.

In Figure 13, the frequency responses of four effective analysis filters of a warped ($\lambda = 0.5$) 64-point-windowed DCT-4 filterbank are plotted. Here the window has been chosen without a special optimization (Kaiser window with $\beta = 10$). Therefore, the capability to approximate gammatone filter frequency responses or roex functions is limited (especially at higher frequencies). Anyway, we can observe that the responses fit relatively well at low center frequencies. Note that this behavior is contrary to what we have observed for the FIR gammatone filter design, where the necessary impulse response length increases with decreasing center frequency. Further optimization of the window will improve the frequency responses.

A window length of only 64 samples yields reasonable frequency responses at a sampling rate of 8 kHz. Therefore, the usage of a frequency-warped transform filterbank constitutes a computationally highly efficient and memory-saving option for an auditory filterbank implementation on a DSP for real-time applications.

Inversion

A synthesis filterbank can be obtained by generalizing the overlap-and-add procedure which is well known from the inverse short-term Fourier transform in the same way as the

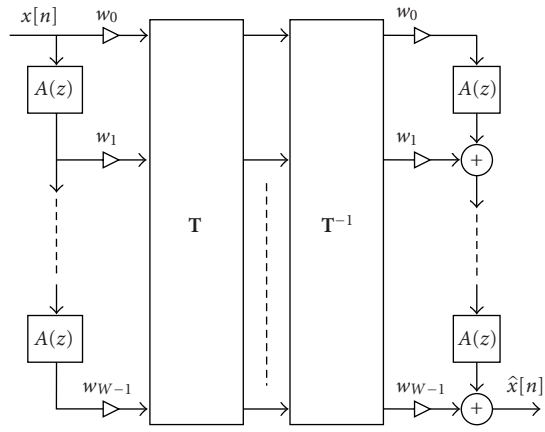

Figure 14: Nondecimated frequency-warped phase-distorted analysis/synthesis filterbank.

sliding window—by replacing the unit-delay chain with a general allpass chain (see Figure 14). While the uniform frequency resolution analysis/synthesis filterbank achieves perfect reconstruction, the frequency-warped version does not. For the simple case, when the window length W equals the number of channels L, we can choose the window coefficients such that $\sum_{i=0}^{W-1} w_i^2 = 1$ and we obtain for the output signal

$$\hat{X}(z) = X(z)A^{W-1}(z), \tag{A.3}$$

and, therefore, a phase distortion is introduced. In [47] an FIR filter is used to compensate for this phase distortion to get a near-perfect-reconstruction filterbank. However, this introduces an additional delay, which increases with decreasing compensation error. Anyway, it is not necessary to equalize for perfect linear phase since small phase distortions are inaudible. The case where a longer prototype filter is used without a higher number of auditory channels, that is, $W > L$, is also considered in [47].

In a recent development [49], we have shown that an FIR synthesis filterbank exists for a critically sampled frequency-warped transform filterbank which achieves perfect reconstruction. However, these synthesis filters amplify any quantization noise introduced in the subband signals and do not exhibit bandpass characteristics. Thus, they are not recommended for coding applications.

ACKNOWLEDGMENT

This paper is an extended version of a plenary lecture presented at the second IEEE Benelux Signal Processing Symposium (SPS-2000) in Hilvarenbeek, The Netherlands, March 2000.

REFERENCES

[1] K. Brandenburg and G. Stoll, "ISO-MPEG-1 Audio: a generic standard for coding of high-quality digital audio," *Journal of the Audio Engineering Society*, vol. 42, no. 10, pp. 780–792, 1994.

[2] B. Tang, A. Shen, A. Alwan, and G. Pottie, "A perceptually based embedded subband coder," *IEEE Trans. Speech Audio Processing*, vol. 5, no. 2, pp. 131–140, 1997.

[3] R. Veldhuis and A. Kohlrausch, "Waveform coding and auditory masking," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds., pp. 427–428, Elsevier Science, Amsterdam, The Netherlands, 1995.

[4] T. Dau, D. Püschel, and A. Kohlrausch, "A quantitative model of the 'effective' signal processing in the auditory system. I. Model structure," *Journal of the Acoustical Society of America*, vol. 99, no. 6, pp. 3615–3622, 1996.

[5] E. Zwicker, "Dependence of post-masking on masker duration and its relation to temporal effects in loudness," *Journal of the Acoustical Society of America*, vol. 75, no. 1, pp. 219–223, 1984.

[6] R. Geiger, A. Herre, G. Schuller, and T. Sporer, "Fine grain scalable perceptual and lossless audio coding based on Int-MDCT," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '03)*, vol. 5, pp. 445–448, Hong Kong, China, April 2003.

[7] M. Hansen and B. Kollmeier, "Using a quantitative psychoacoustical signal representation for objective speech quality measurement," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '97)*, vol. 2, pp. 1387–1390, Munich, Germany, April 1997.

[8] H. Su and P. Mermelstein, "Delayed decision coding of pitch and innovation signals in code-excited linear prediction coding of speech," in *Speech and Audio Coding for Wireless and Network Applications*, B. S. Atal, V. Cuperman, and A. Gersho, Eds., pp. 69–76, Kluwer Academic Publishers, Boston, Mass, USA, 1993.

[9] R. Fandos Marin, *Delayed decision CELP speech coding using squared and perceptual error criteria*, M.S. thesis, Department of Signals, Sensors and Systems, KTH (Royal Institute of Technology), Stockholm, Sweden, 2003.

[10] G. Kubin and W. B. Kleijn, "On speech coding in a perceptual domain," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '99)*, vol. 1, pp. 205–208, Phoenix, Ariz, USA, March 1999.

[11] J. B. Allen, "Cochlear modeling," *IEEE ASSP Mag.*, vol. 2, no. 1, pp. 3–29, 1985.

[12] S. Greenberg, "Acoustic transduction in the auditory periphery," *Journal of Phonetics*, vol. 16, pp. 3–17, 1988.

[13] M. A. Ruggero, "Physiology and coding of sound in the auditory nerve," in *The Mammalian Auditory Pathway: Neurophysiology*, A. Popper and R. Fay, Eds., pp. 34–93, Springer-Verlag, New York, NY, USA, 1992.

[14] E. Zwicker and H. Fastl, *Psychoacoustics. Facts and Models*, Springer-Verlag, Berlin, Germany, 2nd edition, 1999.

[15] B. R. Glasberg and B. C. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, no. 1-2, pp. 103–138, 1990.

[16] B. C. Moore, *An Introduction to the Psychology of Hearing*, Academic Press, London, UK, 4th edition, 1997.

[17] D. D. Greenwood, "A cochlear frequency-position function for several species—29 years later," *Journal of the Acoustical Society of America*, vol. 87, no. 6, pp. 2592–2605, 1990.

[18] A. Härmä, M. Karjalainen, L. Savioja, V. Välimäki, U. K. Laine, and J. Huopaniemi, "Frequency-warped signal processing for audio applications," *Journal of the Audio Engineering Society*, vol. 48, no. 11, pp. 1011–1029, 2000.

[19] R. D. Patterson, I. Nimmo-Smith, D. L. Weber, and R. Milroy, "The deterioration of hearing with age: Frequency selectivity, the critical ratio, the audiogram, and speech threshold," *Journal of the Acoustical Society of America*, vol. 72, no. 6, pp. 1788–1803, 1982.

[20] R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand, "Complex sounds and auditory images," in *Auditory Physiology and Perception*, Y. Cazals, L. Demany, and K. Horner, Eds., pp. 429–446, Pergamon Press, Oxford, UK, 1992.

[21] P. Dallos, "Overview: Cochlear neurobiology," in *The Cochlea*, P. Dallos, A. Popper, and R. Fay, Eds., vol. 8, pp. 1–43, Springer Verlag, New York, NY, USA, 1996.

[22] T. Irino and R. D. Patterson, "A time-domain, level-dependent auditory filter: The gammachirp," *Journal of the Acoustical Society of America*, vol. 101, no. 1, pp. 412–419, 1997.

[23] R. F. Lyon, "A computational model of filtering, detection, and compression in the cochlea," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '82)*, vol. 7, pp. 1282–1285, Paris, France, May 1982.

[24] S. Seneff, "A joint synchrony/mean-rate model of auditory speech processing," *Journal of Phonetics*, vol. 16, pp. 55–76, 1988.

[25] W. Maass and C. M. Bishop, Eds., *Pulsed Neural Networks*, MIT Press, Cambridge, Mass, USA, 1999.

[26] M. Weintraub, "The GRASP sound separation system," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '84)*, vol. 9, pp. 69–72, San Diego, Calif, USA, March 1984.

[27] R. D. Patterson, "A pulse ribbon model of monaural phase perception," *Journal of the Acoustical Society of America*, vol. 82, no. 5, pp. 1560–1586, 1987.

[28] M. Weintraub, "A computational model for separating two simultaneous talkers," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '86)*, vol. 11, pp. 81–86, Tokyo, Japan, April 1986.

[29] M. Slaney, "Pattern playback from 1950 to 1995," in *Proc. IEEE Systems Man Cybern. Conf.*, vol. 4, pp. 3519–3524, Vancouver, BC, Canada, October 1995.

[30] F. S. Cooper, "Acoustics in human communication: Evolving ideas about the nature of speech," *Journal of the Acoustical Society of America*, vol. 68, no. 1, pp. 18–21, 1980.

[31] T. Irino and H. Kawahara, "Signal reconstruction from modified auditory wavelet transform," *IEEE Trans. Signal Processing*, vol. 41, no. 12, pp. 3549–3554, 1993.

[32] M. Slaney, D. Naar, and R. F. Lyon, "Auditory model inversion for sound separation," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '94)*, vol. 2, pp. 77–80, Adelaide, Australia, April 1994.

[33] R. W. Hukin and R. I. Damper, "Testing an auditory model by resynthesis," in *Proc. 8th European Conference on Speech Communication and Technology (EUROSPEECH '89)*, vol. 1, pp. 243–246, Paris, France, September 1989.

[34] X. Yang, K. Wang, and S. A. Shamma, "Auditory representations of acoustic signals," *IEEE Trans. Inform. Theory*, vol. 38, no. 2, pp. 824–839, 1992.

[35] G. Kubin and W. B. Kleijn, "Multiple-description coding (MDC) of speech with an invertible auditory model," in *Proc. IEEE Speech Coding Workshop*, pp. 81–83, Porvoo, Finland, June 1999.

[36] H. Bölcskei, F. Hlawatsch, and H. Feichtinger, "Frame-theoretic analysis of oversampled filter banks," *IEEE Trans. Signal Processing*, vol. 46, no. 12, pp. 3256–3268, 1998.

[37] H. N. Razafinjatovo, "Iterative reconstructions in irregular sampling with derivatives," *J. Fourier Anal. Appl.*, vol. 1, no. 3, pp. 281–295, 1995.

[38] B. Foster and C. Herley, "Exact reconstruction from periodic nonuniform samples," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '95)*, vol. 2, pp. 1452–1455, Detroit, Mich, USA, May 1995.

[39] E. Ambikairajah, J. Epps, and L. Lin, "Wideband speech and audio coding using gammatone filter banks," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '01)*, vol. 2, pp. 773–776, Salt Lake City, Utah, USA, May 2001.

[40] M. Stocker, *Efficient coding methods for a perceptual speech coder*, M.S. thesis, Institute of Communications and Wave Propagation, Graz University of Technology, Graz, Austria, 2003.

[41] C. Feldbauer and G. Kubin, "How sparse can we make the auditory representation of speech?" in *Proc. 8th International Conference on Spoken Language Processing (ICSLP '04)*, Jeju Island, Korea, October 2004.

[42] Y. Wang, "Multiple description coding using non-hierachical signal decomposition," in *Proc. European Conference Signal Processing (EUSIPCO '98)*, pp. 233–236, Rhodes, Greece, September 1998.

[43] M. Slaney, "An efficient implementation of the Patterson-Holdsworth auditory filter bank," Tech. Rep. 35, Apple Computer, New York, NY, USA, 1993.

[44] L. Lin, W. Holmes, and E. Ambikairajah, "Auditory filter bank inversion," in *Proc. IEEE Int. Symp. Circuits and Systems (ISCAS '01)*, vol. 2, pp. 537–540, Sydney, Australia, May 2001.

[45] A. Oppenheim, D. Johnson, and K. Steiglitz, "Computation of spectra with unequal resolution using the fast Fourier transform," *Proc. IEEE*, vol. 59, no. 2, pp. 299–301, 1971.

[46] P. Vary, "Ein Beitrag zur kurzzeitspektralanalyse mit digitalen systemen," Ausgewählte Arbeiten über Nachrichtnesysteme 32, Universität Erlangen, Erlangen, Germay, 1978.

[47] E. Galijašević, "Design of allpass-based non-uniform oversampled DFT filter banks," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '02)*, vol. 2, pp. 1181–1184, Orlando, Fla, USA, May 2002.

[48] J. Smith and J. Abel, "Bark and ERB bilinear transforms," *IEEE Trans. Speech Audio Processing*, vol. 7, no. 6, pp. 697–708, 1999.

[49] C. Feldbauer and G. Kubin, "Critically sampled frequency-warped perfect reconstruction filterbank," in *Proc. European on Circuit Theory and Design Conference (ECCTD '03)*, vol. 3, pp. 109–112, Krakow, Poland, September 2003.

[50] E. Zwicker and E. Terhardt, "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency," *Journal of the Acoustical Society of America*, vol. 68, pp. 1523–1525, 1980.

Christian Feldbauer was born in Grieskirchen, Austria, on May 31, 1976. He received the Dipl.-Ing. degree in electrical engineering/sound engineering from the Graz University of Technology (TUG), Austria, in 2000. The work presented here is part of his Ph.D. research performed at the TUG. Since 2001, he has been a Research and Teaching Assistant at the Signal Processing and Speech Communication Laboratory at the TUG. He was a Guest Researcher at the KTH, Stockholm, in summer 2003 and at the University of Sherbrooke, Canada, in summer 2004. His research interests are in anthropomorphic coding and perception mechanisms for speech and audio, general speech and signal processing, as well as applications and theory of adaptive filters.

Gernot Kubin was born in Vienna, Austria, on June 24, 1960. He received his Dipl.-Ing. (1982) and Dr. Techn. (1990, sub auspiciis praesidentis) degrees in electrical engineering from TU Vienna. He is a Professor of nonlinear signal processing and Head of the Signal Processing and Speech Communication Laboratory (SPSC), TU Graz, Austria, since September 2000. Earlier international appointments include CERN, Geneva, Switzerland (1980); TU Vienna (1983–2000); Erwin Schroedinger Fellow at Philips Natuurkundig Laboratorium, Eindhoven, The Netherlands (1985); AT&T Bell Labs, Murray Hill, USA (1992–1993 and 1995); KTH, Stockholm, Sweden (1998); Vienna Telecommunications Research Centre FTW (Key Researcher and Member of the Board, 1999–now); Global IP Sound, Sweden and USA (Scientific Consultant, 2000–2001); Christian Doppler Laboratory for Nonlinear Signal Processing (Founding Director, 2002–now). He is a Member of the Board of the Austrian Acoustics Association and Vice Chair for the European COST Action 277, Nonlinear Speech Processing. He has authored or coauthored over ninety peer-reviewed publications and three patents.

W. Bastiaan Kleijn holds a Ph.D. degree in electrical engineering from Delft University of Technology, the Netherlands, a Ph.D. in soil science and an M.S. degree in physics, both from the University of California, and an M.S. degree in electrical engineering from Stanford University. He worked on speech processing at AT&T Bell Laboratories from 1984 to 1996, first in development and later in research. Between 1996 and 1998, he held guest professorships at Delft University of Technology, the Netherlands, Vienna University of Technology, and KTH (Royal Institute of Technology), Stockholm. He is now a Professor at KTH and heads the Sound and Image Processing Laboratory in the School of Electrical Engineering. He is also a founder and former Chairman of Global IP Sound where he remains a Chief Scientist. He is on the Editorial Boards of IEEE Signal Processing Letters and IEEE Signal Processing Magazine and held similar positions at the IEEE Transactions on Speech and Audio Processing and the EURASIP Journal on Applied Signal Processing. He has been a Member of several IEEE technical committees, and a Technical Chair of ICASSP-99, the 1997 and 1999 IEEE Speech Coding Workshops, and a General Chair of the 1999 IEEE Signal Processing for Multimedia Workshop. He is a Fellow of the IEEE.